Avoiding a Data Science Winter by Keeping the Expectations Low

Matthias Hofstetter (✉), Adamantios Koumpis
Institut Digital Enabling, Berner Fachhochschule, Bern, Switzerland
matthias.hofstetter@bfh.ch

Kyriakos Chatzidimitriou
Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract—In this paper we present and discuss some aspects related to what we consider as some of the most important corporate challenges of Data Science, AI and Machine Learning regarding both human talents and business. We examine the case of a discussion that took place over Quora and in particular we focus on an answer we have selected as indicative of a potentially threatening situation for the sustainable development of the Data Science, AI and Machine Learning disciplines as well as the growth of the respective demand and supply sides and the corresponding ecosystem these form. We then make an attempt to examine the setting by means of analyzing the case, using as our guide the provided narrative.

Keywords—Data Science, Artificial Intelligence, Machine Learning

1 Introduction

The term AI winter has not been used in a while and many people, not just laypersons, but professionals as well, are not familiar with it. It is actually supposed to denote two periods spanning from the mid-70s until the start of the 80s and then again from the end of the 80s until almost the middle of the 90s. Both these periods were characterized by low investments and expectations from what AI can actually deliver, and as one may easily infer, they followed periods of high optimism and unsparing expectations. Both these winters destroyed much but—same as the winter season in the real world—catered to the springtime and the summer of AI we are now experiencing.

In the following sections we shall present and discuss some aspects related to what we consider as some of the most important corporate challenges of Data Science, AI and Machine Learning regarding both human talents and business.

In contrast to the first AI winter that was supposedly triggered by one person and their report [8], the dangers and threats now may come from a plethora of ignition sources, with a primary position related to the social media and networks.

One watches the global developments of the coronavirus pandemic and grows concerned for what will happen next. Sometimes the past teaches us more about the

future, so there may be some good reasons to 'read the classics' than expect to predict the future by reading the crystal ball of big data and artificial intelligence.

In 1995, Dornbusch, Goldfajn and Valdés published a paper that for some economists holds a special place in the altar of theoretical works in the discipline of economics that emerge from practice and have direct relevance and value for uptake by practitioners in the daily battlefields of the economy.

Their paper entitled, 'Currency Crises and Collapses', inspired a term that is now used widely in economics: *sudden stop* [1]. Though not quite the case, one may regard as a sudden stop a sudden slowdown in capital flows into a market economy. Such a stop is expected to be followed by decreases of outputs, even steeper decreases in private spending and credit to the private sector.

At that time the three economists were sharing their thoughts and ideas about the 1994 economic crisis in Mexico. The term *sudden stop* had been attributed to an anonymous banker's adage, according to whom, "it is not speed that kills, it is the sudden stop".

It might be good to see how Mexico's, and at a greater extent also the United States', economy was saved (answer: by a generous and unimaginably huge and high risk bailout signed by then President Clinton) and try to find similarities with the current crisis. However, today we would like to focus on this anonymous banker's adage: *it is not speed that kills, it is the sudden stop.*

So, each time each of us hears, reads or thinks about the dramatic consequences and the potentially tragic implications that the coronavirus outbreak shall have on global, national or individual economies, it is good to think about the banker's wisdom: *it is not speed that kills, it is the sudden stop.*

Since 1994, things have changed. But in this hyper-optimized, globalized economy the quote looks even more suited. With many employees living paycheck to paycheck, SMEs month by month and large corps quarter to quarter, a sudden stop equals a sudden "death". There is a high possibility that the coronavirus pandemic will not be the end of worldwide crisis scenarios. Thus, at the expense of producing less, winning less, travelling less and probably working more, we should strive to create more robust and distributed clusters of supply-demand chains and value-oriented economies that will defend better against future sudden stops.

In the following sections we examine the case of a discussion that took place over Quora and in particular we focus on an answer we have selected as indicative of a potentially threatening situation for the sustainable development of the Data Science, AI and Machine Learning disciplines as well as the growth of the respective demand and supply sides and the corresponding ecosystem these form. We then make an attempt to examine the setting by means of analyzing the case, using as our guide the provided narrative. Regarding the use of narratives and user stories may seem as emerging and its proliferation is not considered as new at all [5] [9]. However, in the recent years there is a clear trend for bringing it closer to the practice of both technology development and technology use [3].

2   Setting The Stage: "Have You Ever Seen A Data Scientist Get Fired? If Yes, Why?"

In an article that appeared in the "Towards Data Science" publication of the Medium online publishing platform, the author discusses how we usually tend to overestimate the role of hard skills as exemplified by Data Science projects during the job interview process and elaborates on "softer" skills that actually matter [4]. However, and as we all know, getting a job is one (very important) thing and thriving at it is quite different, sometimes a lot tougher.

In science, many times, a good research question matters more than its answers. Answers may (and are usually) wrong—while a question, even one that seems naive, waits sometimes for years to get a good or correct answer.

A relatively recent question on Quora, an international online question-and-answer platform, that sparked our interest was: "Have you ever seen a data scientist get fired? If yes, why?" Below we use the opportunity to share with you some thoughts that we started to develop when we read it and especially when we read the top-voted answer it got.

The top answer started with an unexpected crescendo that one may dare to compare with several great opening paragraphs in epic fiction works namely this here: *'I am actually about to get fired after about 6 months on the job.'* [13]. So, to us it seemed like the start of a thrillingly good coincidence: to find an all-worthy-to-ask question, answered by someone who offers their own bitter experience on the subject.

The respondent offered the following reasons for their soon-to-happen firing from the job *(to be noted that all excerpts used are presented in italics and all come from the aforementioned answer in the Quora platform)*:

'Machine learning is not completely usable to the level that is hyped up to be in every single industry out there. The thinking today is, you toss a bunch of features into a model and the model will predict with 95% accuracy. Here is the truth.

- *Most companies and industries are not ready for ML. They don't have the right set of features, have access to the right set of features or even know what features drive their market.*
- *ML is seen as a magic bullet that can solve anything, which is simply not true. If your feature space is not correlated in some way to the target variable, then no model can have any meaningful predictive power. For example, having a feature such as how many presidents over 50 have won a second term election will be useless to predict who is going to win in 2020. The reason I say this is because those are the types of features that are mostly available, related but not really driving the model.*
- *Most of the problems that companies want to throw at ML are usually not easy problems. They are usually what we call pie in the sky problems like predicting the stock market.*

What is described above is the case of an industry struggling to get along with a new technology, finding themselves in the role of early adopters as the well-known Rogers'

bell curve shows [12]. On the other hand, no one would describe the use of Machine Learning (ML) as innovative today. Attempts to use ML in the 1990s might have been regarded as such—but even at that time there were also still some prior attempts.

Our aim is *not* to deconstruct the arguments made by [13]—quite the opposite—we would like to provide possible explanations for the situation, which undoubtedly may face serious improvements.

2.1 Argument number 1: Blame the companies (and their features)

'Most companies and industries are not ready for ML. They don't have the right set of features, have access to the right set of features or even know what features drive their market.'

Of course, some companies are in a business field that might be calling more for ML applications and a more intensive use of ML, but this is something to be proven. If a company is data-intensive or data-greedy, produces bulk amounts of data and has employees exploiting the data (features) to produce decisions (labels), then this company might appear as a good candidate for ML.

However, we understand that what is meant above is rather the same as for many people who visit a well-being session, or a Mandala-Thai massage, use the services of a psychoanalyst or attend a mass in church: they are all driven by some internal (esoteric) need which they try to satisfy. This is a well-understood case for humans, which usually brings us to the discovery of new knowledge and helps us innovate thus we see no good reason to blame the companies for not being 'of the right type' at all.

Of course, **companies carry responsibilities too**—to name only a few:

- They open positions with no care or consideration on whether they are really capable to support them. Even for some simple data storage expenses, mind calculation expenses, cloud costs might lead a startup to a death spiral (for example see on the costs of training state of the art models [14]). We've seen Small and Medium-sized Enterprises that want to "do Machine Learning" (and the other hyped terms), but once they start calculating the cloud costs for data and computations they cut back.
- There is no lead to connect the individual contributors with the business—one needs a person who can speak both languages and bridge the (many) communication gaps.

They are (yet) unable to see skills other than the hard skills, and are hence not in a position to judge who are the individual contributors, who are the leaders, the product managers, etc. Again [4] covers this aspect sufficiently.

2.2 Argument number 2: The magic bullet

'ML is seen as a magic bullet that can solve anything, which is simply not true. If your feature space is not correlated in some way to the target variable, then no model can have any meaningful predictive power. For example, having a feature such as how

many presidents over 50 have won a second term election will be useless to predict who is going to win in 2020. The reason I say this is because those are the types of features that are mostly available, related but not really driving the model.'

Imagine that somehow a fictional crawler finds out information that has never been considered before: for example, the hypothetical case of all presidents having had as children a tomcat which they all called Bob. This might be a feature we would like to take into account especially if for the next elections there is also one or more candidates with the same feature.

We think that as a data scientist one should worry more about the meaningfulness of the models as such and only at a secondary level about their predictive power.

So again, the aim of companies is mainly business related and ***technology is only a means to this end***. Companies may or may not curate lots of data relevant to their business and may or may not connect these data with other events that take place 'out there'. How successful they are in predicting the future or improving their own business depends on lots of parameters. We see the demand for them to predict the future and come up with killer applications or products. Every failure in business has been prepared or foreseen as a possible or potential success—there is no doubt about this. Sometimes companies, same as people, get blinded and see only one side of a reality. But exactly because of this one-sidedness they also sometimes succeed in their business. We can imagine two-handed political advisors of a candidate for the presidency advising their client to not run for president as they didn't have as a child any tomcat called Bob. This is not science—this is rather occultism. And ML is not a pillar to support the supremacy of the new Dark Ages in the corporate decision-making practices, where data-specialised witches prepare ML-supported magic potions in their cauldrons on a quest to foresee the future.

To conclude: we wish companies to invest in collecting all information relevant to their business and to their market at large, investing generously on modeling exercises which may fail but may also succeed—sometimes one needs a long enough horizon to get good results. It is ***not*** about collecting relevant features and ***not*** about predicting the future. It is mainly about ***understanding the business*** and being able to ***explain the business*** and possibly a little bit of our world. If ML can help us to this, we shall be happy to utilise it as a tool.

2.3   Argument number 3: Apple pies in the sky

[13] provides a third argument to convince us that most companies and industries are not ready for ML, the fact that: *'Most of the problems that companies want to throw at ML are usually not easy problems. They are usually what we call pie in the sky problems like predicting the stock market.'*

There is no doubt that few people might dare to visit a specialist in oncology only for the sake of *generally* discussing cancer treatment options. Usually, people are sent to such a specialist at a very dramatic point in their own lives or in the life of some of their relatives. And what they demand from such a specialist is to offer them the best treatment *'here and now'*. They may not be aware of terms such as personalised or precision medicine, so they wouldn't care to get the same treatment as the previous ten

or 100 patients of the particular specialist, as long as it works, they get healed and are able to enjoy a return to normal life.

Of course, there is always space for selling people snake oil or hope when there is no hope at all. But we don't think that companies are so stupid to grow all the false types of expectations *not in regard to the technology they employ but in regard to the answers they want to get for their own problems*.

We understand very well why [13] names some problems 'pie in the sky'. On the other hand, we don't see any reason why people shouldn't ask for more: predicting the stock market is not an unreasonable aim. It might be something difficult to offer, same as telling people when they are going to die or, better, of what they are going to die, while also giving them some treatment to avoid dying.

So even if this is difficult or impossible, companies—same as people—in their role as customers of information technologies have good reasons to ask for the impossible (and also compensate for the impossible as well). Usually what we understand as impossible now may change in the future. Several impossibilities have been successfully met and answered by science and technology—while several other problems remained unsolvable though a successful answer seemed only some little time away.

3 Mistakes in the Procedure

It is a pity that the story has a starting point that someone is/was about to get fired. But let's see each of the project assignments and try to understand what went wrong. [13] complains that he has: '*been at the job for 6 months and I have been given several ML projects but ultimately none were solvable. lol. This is because every single problem that they want me to solve is a 'pie in the sky' problem.*'

We understand that feeling and have several times been confronted with it—it is certainly one that kills a person's morale. Universities educate people to get degrees—so they pass through certification procedures, where the most talented or the ones that have been better prepared get the best scores (on this see also [2] for an elaboration on the matter), but in the real world as a stage we are all desperately unrehearsed.

According to [13] for his first project he had to: '*reverse engineer a famous tech company's algorithm and outsmart their algorithm and essentially game the system. This was doomed to fail from the beginning. Now mind you, this is a famous company that brings in millions + in profit using this algorithm. The point was to outsmart this algorithm with my ML efforts so we don't have to pay this company as much money. I'm supposed to find an ML loophole that has predictive power. Or better yet, this Fortune 500 company is so stupid that anyone could crack their algorithm! Yeah right! Ultimately, this project was unsuccessful with the predictive power of a random guess.*'

One may very well imagine the expectations of the people who assigned this task to someone. They envisioned that by hiring the person, they were entering the motorway of Data Science, something that all presentations and articles advertise, where little Davids with a DIY slingshot may compete and win out over Goliaths. There is nothing wrong about this vision. Anyone's task as an employee in a company that would ask them something like this should be to slice the vision into many smaller parts (so exactly

what [4] refers to as *analytical thinking*) and try to make progress every day or week that can also be reported and communicated to the employer so that they a) wouldn't think one does nothing, and b) have arguments for every decision made, since 'code wins arguments' [10]. Such a communication might offer the opportunity to get feedback. *Perhaps* the way one slices the vision into smaller entities has been wrong. *Perhaps* one is already lost and needs some help or some more elaborate directions.

The second project was '*[a]nother problem like the above but this time trying to predict the future of the company in one way or another. This was partially successful.*'

Success in some cases is random. On the other hand, and for that (partially) successful case, the most important outcome would not have been the accuracy of the model per se, but to reverse engineer the process and find the reason why the first problem (and the third as we shall see later on) was a failure, while the second one, which according to [13] was similar to the first, was a success. Such a feedback loop (or retrospective) is a very important step in the long term. In addition, generating and communicating short-term wins is also one of the steps leading and driving change in an organization according to Kotter's 8-Step Process for Leading Change [7], in this case towards enabling ML.

Last but not least, the third project was about: '*Predict company profit at the site level. You can already see where this is going. This company has 1000s of sites across the country, akin to a retail chain. I'm supposed to come up with one model that captures the behavior to predict the profit for every site at the site level! We are actually able to get a lot of census statistics to make sense of this. For example, you'd think all you'd need are features like population, jobs, etc., to see how this will drive the market. Again, this is like trying to predict the stock market because common census statistics alone do not drive profit. We have every feature we could think of but still a single model with a set of features is unable to capture the behavior of say New York vs Modesto convincingly because ultimately it is not just about job growth or the people that live there but there are other hidden socioeconomic variables that we cannot capture or we don't have access to or even know about. This model actually produced reasonable predictive power but they want 90%+ accuracy so the project is deemed a failure.*'

From our own personal experiences that we have acquired in teaching students who had been taking before introductory as well as advanced courses on Data Science, the students are trained to use some tool and/or programming language and cope with assignments which are not very complicated and are in no case comparable to the ones that someone faces in real life. We admit that an extra difficulty of real life is that people usually set the bar high as far as their expectations are concerned. So, while a student has been educated to deal with elementary arithmetic problems in the class as primary school pupils do, in real life they need to adapt to the needs and the demands of high-school or college algebra, while sometimes the customers or their bosses demand from them a level comparable to a top-tier university professor. Again, the answer on how to cope with such a situation has been provided in [4] and mentioned as *analytical thinking*: one breaks the problem into subproblems, discusses them with other people and tries to build a stable base of knowledge. This is the real asset of the company and to this it is where the person or persons are supposed to contribute and not to systems that exhibit fragile accuracies.

A fear one may develop by reading other similar stories is that employers may demonise ML in general as a result of their failure to chart this terra incognita that ML is for companies today. The main problem with expectations from the side of an employer or a customer is the overpromising that takes place from the side of the seller or the employee. And it is a bitter truth that many data scientists sell their projects and not their experience in solving problems, because this is probably what companies and hiring managers want to hear about.

Quoting John Kavanagh who famously keeps saying "Win or Learn" [6], it also applies to both data scientists and ML experts as employees as well as to employers who plan to start their first or next ML project.

4   Conclusions

One may consider the limitations of a technology as related to the technology itself, but most of the time it relates mainly to our own ways of deploying it. Data Science, AI and Machine Learning have been approached as a panacea to a wide range of problems and needs in the corporate world. However, many of the problems they are expected to solve need a framing that we shouldn't expect to be offered for free—it is still our part to carry out the modelling part—otherwise, one shall end up having an ill-defined problem that will not lead to a successful outcome.

Many of the problems that companies face nowadays and which they consider as related to Big Data are exactly the opposite: they are cases of small data. However, if one takes a closer look, they will recognize that it is not as simple as one may think—they are *rich small data*. This means that they hide too much information that expects to be discovered and valuated to set in motion a broad range of corporate decision-making chain reactions.

In the example case we presented in this paper we tried to approach a problem that we see as representative in today's business from the perspective of a story that appeared in the Quora platform.

One needs to be very partial to stories about how asking the right question can sometimes spur major scientific advances. Johannes Kepler may more probably be the hero of natural scientists because of his question "Why circles?", leading to a revolution in astrophysics, but this is no excuse for business analysts and practitioners to avoid asking the same question when faced with the dilemma: "Why Data Science? Why AI? Why ML?"

5   Acknowledgement

6   References

[1] Dornbusch Rudiger, Goldfajn Ilan, Valdés Rodrigo O., *Currency Crises and Collapses*, Brookings Papers on Economic Activity, Vol. 1995, Nr. 2, 1995. https://doi.org/10.2307/2534613

[2] Graham Paul (2019) *The lesson to unlearn*, http://www.paulgraham.com/lesson.html , last accessed: June 19 2020.

[3] Hedman, J., Bødker, M., Gimpel, G., & Damsgaard, J. (2019) Translating evolving technology use into user stories: Technology life narratives of consumer technology use, Information Systems Journal, 29(6), 1178-1200. https://doi.org/10.1111/isj.12232

[4] Jung Haebichan (2019) *How we overestimate the role of Data Science projects in the job process — and the 4 interviewing skills that actually matter*, Dec 2, 2019, Medium – Towards Data Science, https://towardsdatascience.com/sorry-projects-dont-get-you-jobs-3e5d8e74bfdc, last accessed: June 19 2020.

[5] Kankainen, A., Vaajakallio, K., Kantola, V., & Mattelmäki, T. (2012) *Storytelling Group– a co-design method for service design,* Behaviour & Information Technology, 31(3), 221-230. https://doi.org/10.1080/0144929x.2011.563794

[6] Kavanagh, John (2016) *Win or Learn*, Penguin.

[7] Kotter John P. (2002) The heart of change: Real-life stories of how people change their organizations, Harvard Business School, Boston, Mass. 2002.

[8] Lighthill James (1973) *Artificial Intelligence: A General Survey*, Science Research Council.

[9] Lucassen, G., Dalpiaz, F., Van Der Werf, J. M. E., & Brinkkemper, S. (2015) *Forging high-quality user stories: towards a discipline for agile requirements*, in2015 IEEE 23rd international requirements engineering conference (RE) (pp. 126-135). IEEE. https://doi.org/10.1109/re.2015.7320415

[10] Meijer Erik (2016) *The Hacker Way talk*, May 31, 2016, https://www.youtube.com/watch?v=2u0sNRO-QKQ, last accessed: June 19 2020.

[11] Riedl, M. O., Bulitko, V. (2013) *Interactive narrative: An intelligent systems approach*, AI Magazine, 34(1), 67-67. https://doi.org/10.1609/aimag.v34i1.2449

[12] Rogers Everett (2003) *Diffusion of Innovations*, 5th Edition, Simon and Schuster.

[13] Singer John (2019) Quora, https://www.quora.com/Have-you-ever-seen-a-data-scientist-get-fired-If-yes-why/answer/John-Singer-59 , #793406064, 4 Aug 2019, last accessed: June 19 2020.

[14] Synced Review (2019) *The Staggering Cost of Training SOTA AI Models*, Jun 27, 2019, https://medium.com/syncedreview/the-staggering-cost-of-training-sota-ai-models-e329e80fa82 , last accessed: June 19 2020.

7   Authors

Matthias Hofstetter is a Professor at the Business Department of the Berner Fachhochschule and member of the Institute Digital Enabling with specialization in Information Management, IT Governance and Business Information Systems.

Adamantios Koumpis is a Professor at the Business Department of the Berner Fachhochschule and member of the Institute Digital Enabling with specialization in Digital Business and Business Information Systems.

Kyriakos Chatzidimitriou is an intelligent systems, data and software engineer and researcher. His interests are in the broad area of Software Engineering and applied Machine Learning. He currently works as a postdoc in the Electrical and Computer Engineering department of the Aristotle University of Thessaloniki in the area energy analytics and decision making through Machine Learning.